Neighborhoods of Hamburg - Which one is for you?

October 9, 2019

Neighborhoods of Hamburg - Which one is for you?

Capstone project for Coursera course Applied Data Science Capstone

By Nima Mehrafshan

Hamburg, 7 October 2019

Contents

1 The Business Problem

Deciding which neighborhood to live in or where to buy property is a difficult task, especially if you don't know the ins and outs of the city you're looking at. There are so many factors that may influence your decision and the factors may be different for each decision maker depending on their goals, tastes, preferences and individual situation. Those factors may include the accessibility of the location, distance and quality of schools, the availability of restaurants and cafés, property prices and expectations of future development, the type of people you typically meet in the area, and so on. In this project, I'd like to explore the 104 neighborhoods of my home town Hamburg (Germany), and see how data could help to make decisions like that.

I will take the position of a home buyer here because I think everyone can relate to their decision situation. But I think the same analysis would also be relevant to an investor or even to the city administration, who could use it to support their urban development decisions.

So the main questions, I will try to answer are: 1. What are the main drivers of real estate prices in Hamburg? 2. Which neighborhoods are undervalued, when relating their attributes to current average property prices? 3. How can neighborhoods be characterized? Can the 100+ neighborhoods be clustered by similarity into a managable number of groups, in order to get a quick understanding of what they are like?

2 Research Outline

1) To answer question one I will estimate a random forest model with property price levels as the target and the socio-economics as well as the points-of-interest as features (hedonic pricing model). I will then extract the importances from the model to identify the main drivers of property prices in Hamburg. The direction of the relationship From can then be derived from the correlations of these features with the price levels.

2) For question two I will predict price levels for all neighborhoods using the model from above and use the residuals as an indicator of under or overvaluation.

3) I will use the relevant features derived in 1) to run a KMeans cluster analysis and then characterize the resulting clusters based on the most prevalent differences of the clusters compared with the city average.

3 Data

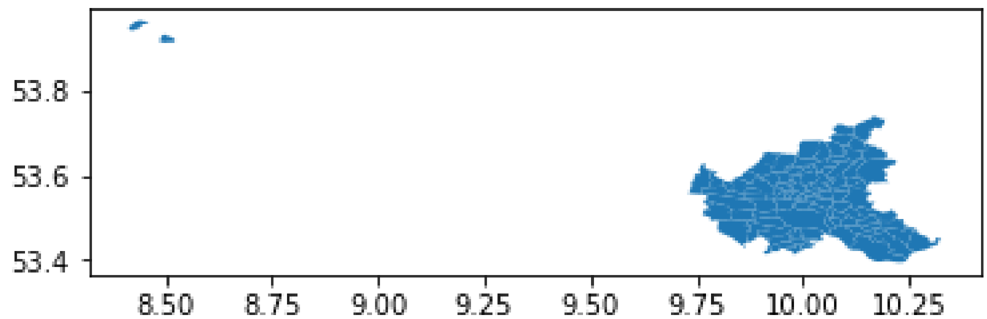
I will need data about average prices in the neighborhoods and about as many neighborhood characteristics as possible. I found three main data sources that I will deploy in the analysis:

- Socio-economic data from Statistikamt Nord, the statistics office for the Bundesländer (German states) Hamburg and Schleswig-Holstein. They provide over 70 socio-economic variables for the years 2013 - 2017, including average household income, number of schools (by type of school), average household size, unemployment rates, and many more. The data set also includes *average property prices per m^2* !
- Points of interest from the Foursquare API, such as restaurants, grocery stoes, nightlife venues, etc.
- The borders of the Hamburg neighborhoods in a digital format from the land surveying office (Landesbetrieb Geoinformation und Vermessung), which I will use for plotting and to link the Foursquare venues to the neighborhoods.

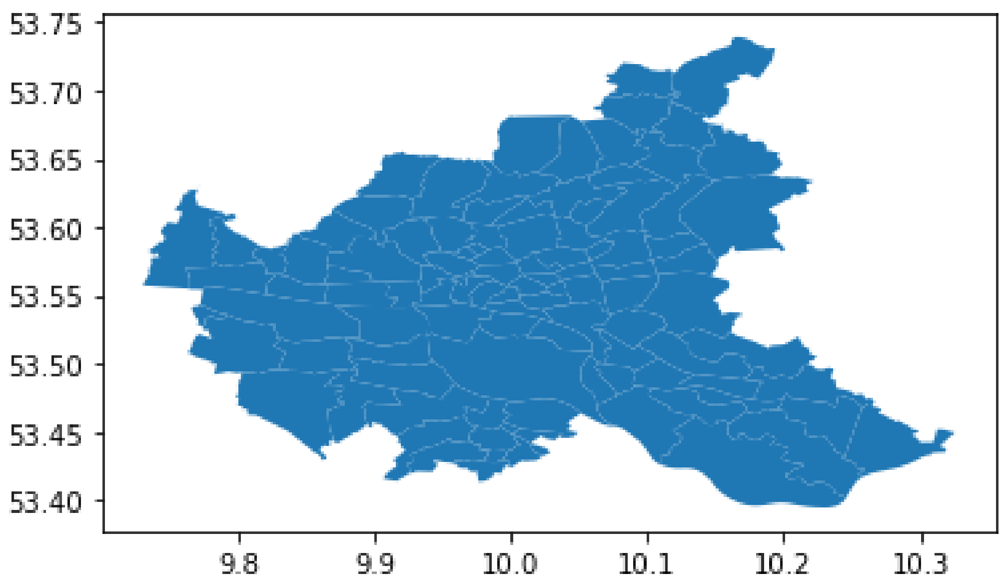
3.1 Data Acquisition

3.1.1 Neighborhood Shapes

To get the neighborhood shapes I am using the WFS service of the land surveying office, transform it into a GeoDataFrame and store the result in a geoJSON file.



The plot shows that the exclave neighborhood "Neuwerk" is included in the data. Even though it legally belongs to Hamburg, it has litte relevance with only ~50 inhabitants, so I am removing it:



3.1.2 Socio-economic Data

Next I get the socio-economic data from the statistics office. It is provided as Excel files. I wrote a function `import_socioecon_data` that takes care of a number of things: - Downloading the files and reading the data - Storing meta data, such as the source variable names and descriptions, into a file - Translating the variable names and descriptions to English using the Google Translator API - Reading in the meta data file if the function is being run a second time and using the variable names that I have edited in the meantime - Merging the data from the different files - Specifying the administrative levels (the data contains data on neighborhoods, city districts and the city as a whole) - Saving the data as a pickle file

Here are the first few lines:

[104]:
```
              name  area_in_km2  average_apartment_price_per_m2  \
0  Hamburg-Altstadt     1.300347                          5710.0
1         HafenCity     2.425612                          7958.0
2          Neustadt     2.261902                          5081.0
3         St. Pauli     2.242488                          5991.0
4         St. Georg     1.822657                          5169.0

   average_house_price_per_m2  average_household_size  average_income  \
0                         NaN                1.592838           31336
1                         NaN                2.115759           93206
2                         NaN                1.498001           34521
3                         NaN                1.532348           27977
4                         NaN                1.515055           44121

   average_land_price_per_m2  average_living_space_m2  births  car_density  \
0                        NaN                73.389302      46   272.885033
1                        NaN                92.797023      68   264.681555
2                        NaN                63.007217     152   246.167152
3                     1316.0                64.234276     239   194.213591
4                     2179.0                71.089805     116   208.050656

   …  students_in_secondary_schools  taxpayers  unemployed_population  \
0  …                             61       1952                     98
1  …                            138       1255                     86
2  …                            400       7015                    504
3  …                            818      11066                   1316
4  …                            299       5683                    427

   unemployed_population_pct  welfare_receivers  welfare_receivers_pct  year  \
0                   5.441421                244              10.585683  2017
1                   3.300077                487              13.427075  2017
2                   5.236364                969               7.618523  2017
3                   7.420355               2987              13.274966  2017
4                   4.909739                819               7.408412  2017

   youth_unemployed  youth_unemployment_pct  admin_level
0               8.0                4.733728           10
1              11.0                3.197674           10
2              26.0                2.546523           10
3              60.0                3.186405           10
```

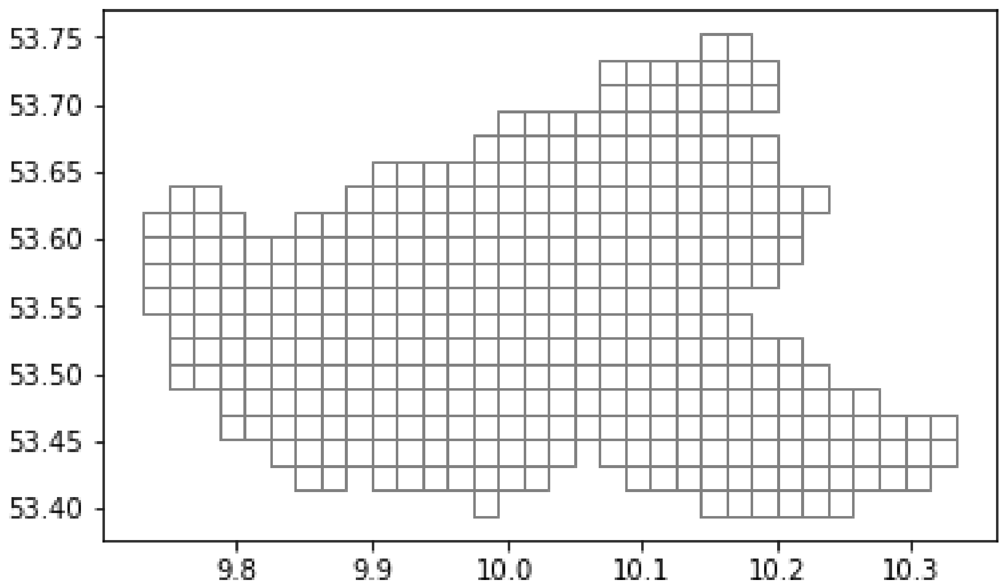
```
4               25.0                    2.149613           10

[5 rows x 79 columns]
```

3.1.3 Points of interest

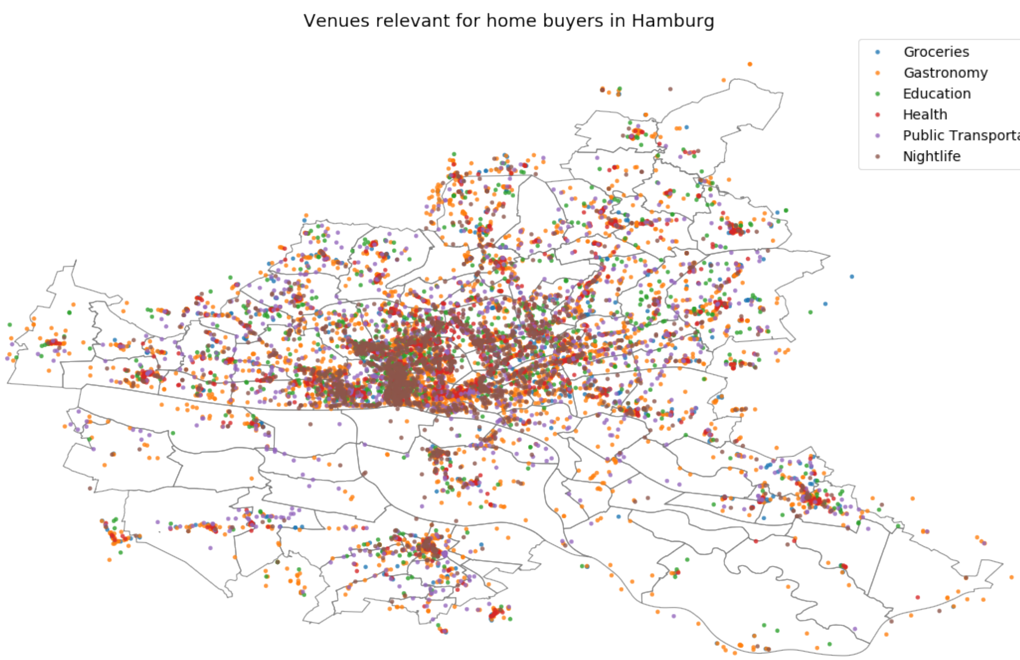
Now to the Foursquare API. I went through the list of Foursquare's venue categories and selected those I thought of being particularly relevant for buyers of apartments or houses. I grouped those categories into six topics:

- Groceries
  - Supermarket
  - Grocery Store
  - Organic Grocery
- Education
  - High School
  - Middle School
  - Child Care Service
  - Playground
- Health
  - Physical Therapist
  - Doctor's Office
- Public Transportation
  - Bus Stop
  - Light Rail Station
  - Metro Station
  - Train Station
- Gastronomy
  - Food
- Nightlife
  - Nightlife Spot

So, I want to get all venues from these categories in Hamburg. Since the API has a limit of 50 results per request, it's not enough to do one request per neighborhood if you don't want to miss out on any venues, because there will of course be more then 50 venues probably in most neighborhoods. Thus, I wrote a function `get_foursquare_data` that loops through smaller raster segments of Hamburg, making a request for each of the topics and if the API returns 50 results, it devides the segment in smaller ones until there are less then 50 results per segment and topic. I start with a raster of 1,000 segments:



After running the function to get the venues let's create a DataFrame with all venues, spatially merge it with the neighborhood shapes, and plot them:



3.2 Data preparation

The socio-economics data is in long format with a row for each year of data for every neighborhood:

[233]:

```
          name  year  area_in_km2  average_apartment_price_per_m2  \
506  Allermöhe  2013         8.65                             NaN
399  Allermöhe  2014         8.65                             NaN
292  Allermöhe  2015         8.65                          2351.0
185  Allermöhe  2016         8.64                             NaN
78   Allermöhe  2017         8.64                          3128.0

     average_apartment_size_m2  average_house_price_per_m2  \
506                     114.71                      2126.0
399                     115.41                      2287.0
292                     115.41                      2270.0
185                     115.19                      2548.0
78                      114.85                      2751.0

     average_household_size  average_income  average_land_price_per_m2  \
506                    2.10           35822                      169.0
399                    2.08           35822                      187.0
292                    2.06           35822                      164.0
185                    2.09           38369                      203.0
78                     2.06           38369                      226.0

     births  …  students_in_comprehensive_schools_pct  \
506      10  …                                  53.90
399       7  …                                  56.18
292       7  …                                  67.03
185       9  …                                  61.45
78       10  …                                  57.14

     students_in_gymnasiums_pct  students_in_secondary_schools  taxpayers  \
506                       44.90                             89        653
399                       40.45                             89        653
292                       30.77                             91        653
185                       36.14                             83        696
78                        41.56                             77        696

     unemployed_population  unemployed_population_pct  welfare_receivers  \
506                     22                       2.40                 29
399                     23                       2.50                 27
292                     25                       2.63                 26
185                     17                       1.84                 34
78                      21                       2.27                 17

     welfare_receivers_pct  youth_unemployed  youth_unemployment_pct
```

```
506                     2.20            3.0                 2.10
399                     2.00            NaN                  NaN
292                     1.88            3.0                 1.64
185                     2.50            0.0                 0.00
78                      1.25            3.0                 2.04

[5 rows x 78 columns]
```

For the following analyses, I'm reshaping it to wide format with one column for each year and variable:

[234]:
```
                 area_in_km2_y2013  area_in_km2_y2014  area_in_km2_y2015  \
name
Allermöhe                     8.65               8.65               8.65
Alsterdorf                    3.06               3.06               3.06
Altengamme                   15.61              15.61              15.61
Altona-Altstadt               2.75               2.75               2.75
Altona-Nord                   2.18               2.18               2.18

                 area_in_km2_y2016  area_in_km2_y2017  \
name
Allermöhe                     8.64               8.64
Alsterdorf                    3.16               3.16
Altengamme                   15.61              15.61
Altona-Altstadt               2.72               2.72
Altona-Nord                   2.22               2.22

                 average_apartment_price_per_m2_y2013  \
name
Allermöhe                                         NaN
Alsterdorf                                     3239.0
Altengamme                                        NaN
Altona-Altstadt                                4022.0
Altona-Nord                                    4022.0

                 average_apartment_price_per_m2_y2014  \
name
Allermöhe                                         NaN
Alsterdorf                                     3747.0
Altengamme                                        NaN
Altona-Altstadt                                4053.0
Altona-Nord                                    4053.0

                 average_apartment_price_per_m2_y2015  \
name
Allermöhe                                      2351.0
Alsterdorf                                     3891.0
```

```
Altengamme                                     NaN
Altona-Altstadt                             4911.0
Altona-Nord                                 4911.0

                 average_apartment_price_per_m2_y2016  \
name
Allermöhe                                      NaN
Alsterdorf                                  4261.0
Altengamme                                     NaN
Altona-Altstadt                             5745.0
Altona-Nord                                 5745.0

                 average_apartment_price_per_m2_y2017  …  \
name                                                   …
Allermöhe                                   3128.0  …
Alsterdorf                                  4486.0  …
Altengamme                                     NaN  …
Altona-Altstadt                             6064.0  …
Altona-Nord                                 6064.0  …

                 youth_unemployed_y2013  youth_unemployed_y2014  \
name
Allermöhe                           3.0                     NaN
Alsterdorf                         18.0                    25.0
Altengamme                          3.0                     8.0
Altona-Altstadt                   102.0                   120.0
Altona-Nord                        79.0                    59.0

                 youth_unemployed_y2015  youth_unemployed_y2016  \
name
Allermöhe                           3.0                     0.0
Alsterdorf                         17.0                    27.0
Altengamme                          4.0                     4.0
Altona-Altstadt                   110.0                   101.0
Altona-Nord                        72.0                    66.0

                 youth_unemployed_y2017  youth_unemployment_pct_y2013  \
name
Allermöhe                           3.0                           2.1
Alsterdorf                         32.0                           1.4
Altengamme                          0.0                           1.3
Altona-Altstadt                   112.0                           4.0
Altona-Nord                        61.0                           3.9

                 youth_unemployment_pct_y2014  youth_unemployment_pct_y2015  \
name
Allermöhe                                 NaN                          1.64
```

```
Alsterdorf                                     1.81                          1.21
Altengamme                                     3.27                          1.68
Altona-Altstadt                                4.68                          4.39
Altona-Nord                                    2.94                          3.58

                youth_unemployment_pct_y2016  youth_unemployment_pct_y2017
name
Allermöhe                               0.00                          2.04
Alsterdorf                              1.82                          2.03
Altengamme                              1.68                          0.00
Altona-Altstadt                         3.98                          4.29
Altona-Nord                             3.29                          2.91

[5 rows x 332 columns]
```

As it turns out, there are some 700+ missing values in the data:

```
Missing values: 739 (0.02%) of 32868
```

I use the iterative imputation of sklearn to predict the missing values based on all available data.

```
Missing values: 0 (0.0%) of 32868
```

Most variables in the dataset are available in absolute terms and as a percentage of some base, e. g. the population under 15 on welfare (`population_under_15_on_welfare`) and the population under 15 on welfare as a percentage of the total population (`population_under_15_on_welfare_pct`). And after the reshape, there is now a column for each year. To simplify the dataset, I will calculate a 5 year change for all absolute variables and then drop them. I will also drop all but the 2017 versions of the variables.

[109]:

```
                area_in_km2  average_apartment_price_per_m2  \
name
Allermöhe          8.637358                     3128.000000
Alsterdorf         3.155241                     4486.000000
Altengamme        15.608172                     2718.152032
Altona-Altstadt    2.717873                     6064.000000
Altona-Nord        2.217817                     6064.000000

                average_apartment_size_m2  average_house_price_per_m2  \
name
Allermöhe                      114.851852                 2751.000000
Alsterdorf                      77.606108                 6451.000000
Altengamme                     107.537954                 2183.000000
Altona-Altstadt                 63.313431                 6094.660447
Altona-Nord                     63.919232                 5912.707922

                average_household_size  average_income  \
name
Allermöhe                     2.059880         38369.0
```

```
Alsterdorf                  1.806265        52426.0
Altengamme                  2.226155        47341.0
Altona-Altstadt             1.652749        30833.0
Altona-Nord                 1.651354        29901.0

                 average_land_price_per_m2  births  car_density  deaths  …  \
name                                                                     …
Allermöhe                            226.0    10.0   559.882439    16.0  …
Alsterdorf                           886.0   168.0   346.550462   157.0  …
Altengamme                           230.0    19.0   565.449688    30.0  …
Altona-Altstadt                     1465.0   390.0   228.101197   211.0  …
Altona-Nord                         1153.0   331.0   225.137279   113.0  …

                 residential_buildings_5y_chg_pct  \
name
Allermöhe                              100.562500
Alsterdorf                             102.207450
Altengamme                             100.740506
Altona-Altstadt                        100.833123
Altona-Nord                            100.845343

                 serviced_apartments_5y_chg_pct  single_households_5y_chg_pct  \
name
Allermöhe                             70.428571                    113.102564
Alsterdorf                           133.408602                    107.285643
Altengamme                            11.500000                    107.360129
Altona-Altstadt                      428.411765                    100.689736
Altona-Nord                          100.000000                    100.590250

                 singleparent_households_5y_chg_pct  \
name
Allermöhe                                 91.307692
Alsterdorf                                99.279330
Altengamme                                86.931034
Altona-Altstadt                           97.913043
Altona-Nord                               99.153846

                 social_housing_units_5y_chg_pct  \
name
Allermöhe                             100.000000
Alsterdorf                             80.440443
Altengamme                             99.000000
Altona-Altstadt                        79.977312
Altona-Nord                            92.212036

                 students_in_secondary_schools_5y_chg_pct  \
name
```

```
Allermöhe                                          85.516854
Alsterdorf                                        103.685212
Altengamme                                         81.035928
Altona-Altstadt                                   105.902502
Altona-Nord                                        99.526316

                 taxpayers_5y_chg_pct  unemployed_population_5y_chg_pct  \
name
Allermöhe                  105.584992                          94.454545
Alsterdorf                  96.569283                         113.369501
Altengamme                  98.624060                          99.000000
Altona-Altstadt            104.151065                          94.781638
Altona-Nord                100.139252                          91.903752

                 welfare_receivers_5y_chg_pct  youth_unemployed_5y_chg_pct
name
Allermöhe                           57.620690                    99.000000
Alsterdorf                         148.662618                   176.777778
Altengamme                          84.000000                    -1.000000
Altona-Altstadt                     90.348343                   108.803922
Altona-Nord                         98.382716                    76.215190

[5 rows x 102 columns]
```

To merge the socio-economic data with the shapes, I can only use the neighborhood names as keys. So, let's first check out if there are any differences in the names:

```
Columns in shapes not in data: Neuland, Steinwerder, Neuwerk, Kleiner Grasbrook,
Finkenwerder, St.Pauli, Altenwerder, St.Georg, Waltershof, Moorburg, Gut Moor
```

```
Columns in data not in shapes: Waltershof und Finkenwerder, Kleiner Grasbrook
und Steinwerder, St. Georg, St. Pauli, Neuland und Gut Moor, Moorburg und
Altenwerder
```

The statistics office apparently has merged some neighborhoods with low populations. Moreover, in the shapes data, there are no spaces after the "." in "St.Georg" and "St. Pauli". So I fix those difference and merge the neighborhoods also in the shapes dataset.

[111]:

```
                                  name       district  \
97                            Wilstorf        Harburg
98   Kleiner Grasbrook und Steinwerder  Hamburg-Mitte
99            Moorburg und Altenwerder        Harburg
100               Neuland und Gut Moor        Harburg
101        Waltershof und Finkenwerder  Hamburg-Mitte

                                              geometry
97   MULTIPOLYGON (((9.97677 53.44140, 9.97640 53.4…
98   MULTIPOLYGON (((9.97670 53.52761, 9.97669 53.5…
99   MULTIPOLYGON (((9.88750 53.49915, 9.88750 53.4…
```

```
100  MULTIPOLYGON (((10.00506 53.47249, 10.00651 53…
101  MULTIPOLYGON (((9.90308 53.54164, 9.91310 53.5…
```

Now we can go ahead and merge the two datasets:

Next, we use the venues data to create a DataFrame containing the count of venues of each *topic* (see above) per neighborhood.

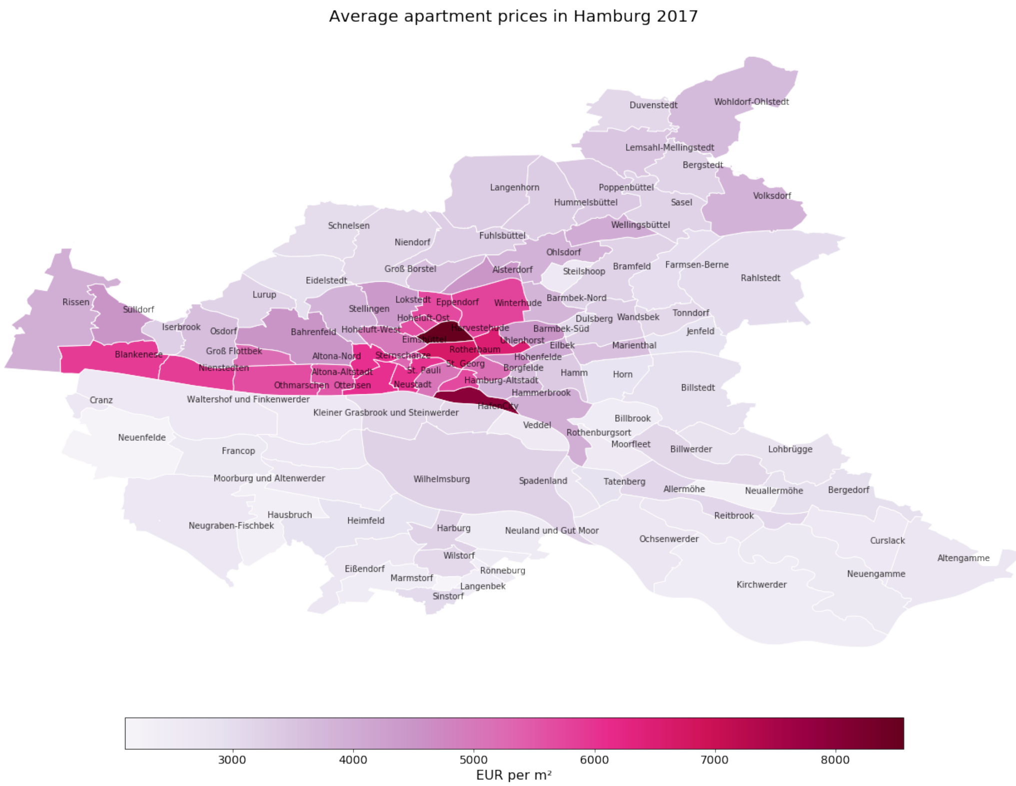
[113]:
```
                  education_venues  gastronomy_venues  groceries_venues  \
name
Allermöhe                      0.0                2.0               0.0
Alsterdorf                     7.0               29.0               6.0
Altengamme                     0.0                4.0               0.0
Altona-Altstadt               20.0               64.0              20.0
Altona-Nord                   16.0               45.0              12.0

                  health_venues  nightlife_venues  public_transportation_venues
name
Allermöhe                   0.0               2.0                           1.0
Alsterdorf                 13.0               5.0                          11.0
Altengamme                  1.0               1.0                           0.0
Altona-Altstadt            62.0              41.0                          28.0
Altona-Nord                15.0              25.0                          76.0
```

Finally, we merge these venue counts with the socio-economics data and generade a table of summary statistics (only first and last 5 columns here for brevity).

[63]:
```
                              count     mean      std      min      25%  \
area_in_km2                    99.0     7.79     7.00     0.55     2.73
average_apartment_price_per_m2 99.0  3762.55  1322.55  2107.00  2805.00
average_apartment_size_m2      99.0    84.26    19.89    50.59    69.84
average_house_price_per_m2     99.0  4083.90  1513.93  2107.00  2793.84
average_household_size         99.0     1.88     0.26     1.30     1.66
…                               …       …        …        …        …
gastronomy_venues              99.0    49.25    49.93     0.00    12.50
groceries_venues               99.0     8.43     7.61     0.00     2.50
health_venues                  99.0    15.26    17.30     0.00     3.00
nightlife_venues               99.0    19.62    32.21     0.00     2.00
public_transportation_venues   99.0    13.30    12.21     0.00     4.50

                                   50%      75%      max
area_in_km2                       5.95     9.95    35.39
average_apartment_price_per_m2 3265.00  4494.00  8560.00
average_apartment_size_m2        78.42    98.53   143.60
average_house_price_per_m2     3696.00  5192.30  8416.00
average_household_size            1.89     2.09     2.73
…                                 …        …        …
gastronomy_venues                33.00    62.00   212.00
```

```
groceries_venues                      7.00     12.50     33.00
health_venues                         9.00     22.50     73.00
nightlife_venues                      8.00     23.50    220.00
public_transportation_venues         11.00     17.00     76.00

[108 rows x 8 columns]
```

The following map shows prices per m² for apartments accross the neighborhoods of Hamburg.



4 Methodology and Analysis

In this section I will conduct the actual analysis. I start with deriving the main drivers of property prices.

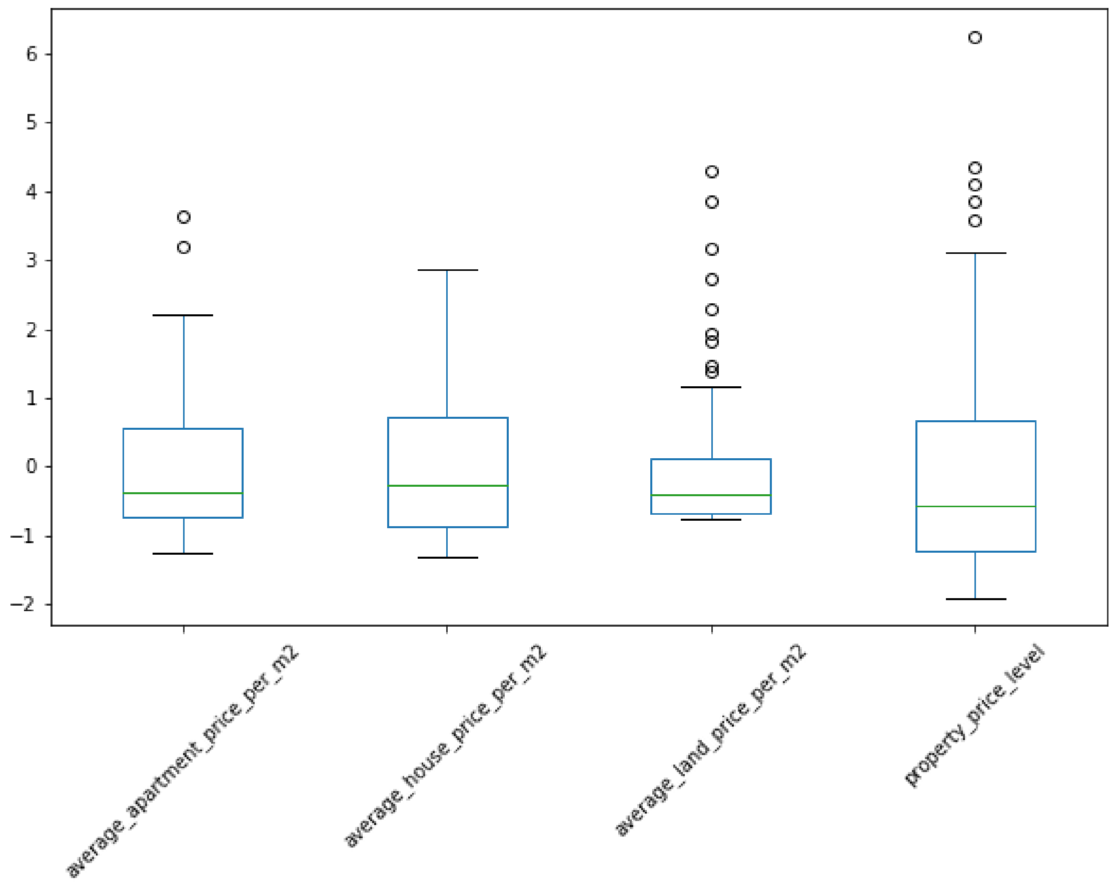
4.1 Main drivers of property prices

The data contains three property price variables: - Average price per m² of apartments (condos) - Average price per m² of houses (detached and semi-detached) - Average price per m² of building land

In some neighborhoods one type of property is more prevalent than the other. In order to have a single target variable, I will reduce these three variables to one using Principal Component Analysis (PCA) and call it *property price level*. Before doing that I will standardize the data ($\mu = 0$, $\sigma = 1$).

nb. I am dropping the area of the neighborhood because it is not a very actionable variable.

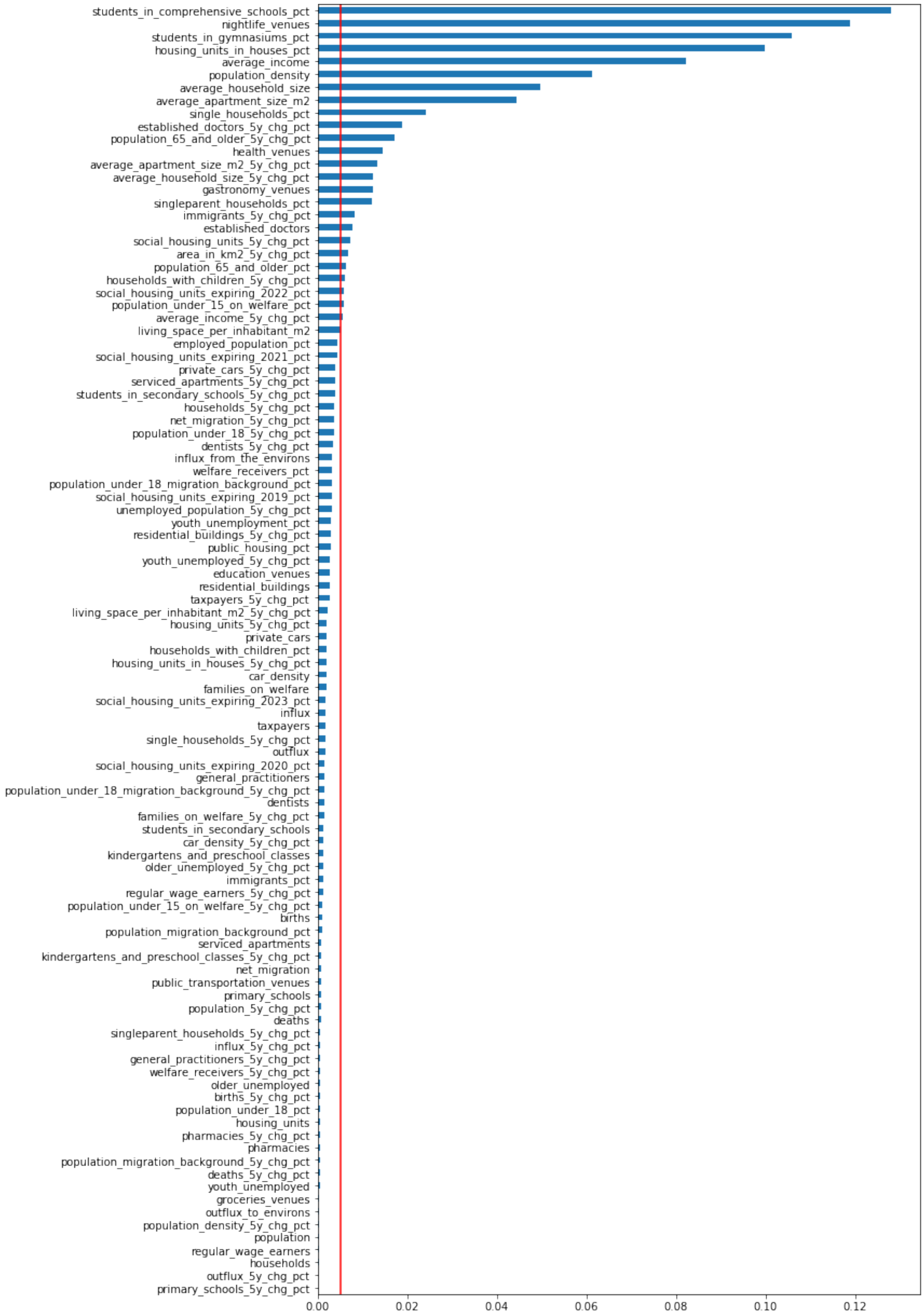
Here is a histogram that shows the distribution of the three standardized variables and the principal component I called property price level.



Now, let's estimate a random forest model to derive feature importances. I will use 5-fold cross validation to tune the main parameters of the models and use the *Mean Squared Error* as the cost function.

```
The model's R² is: 0.9532
```

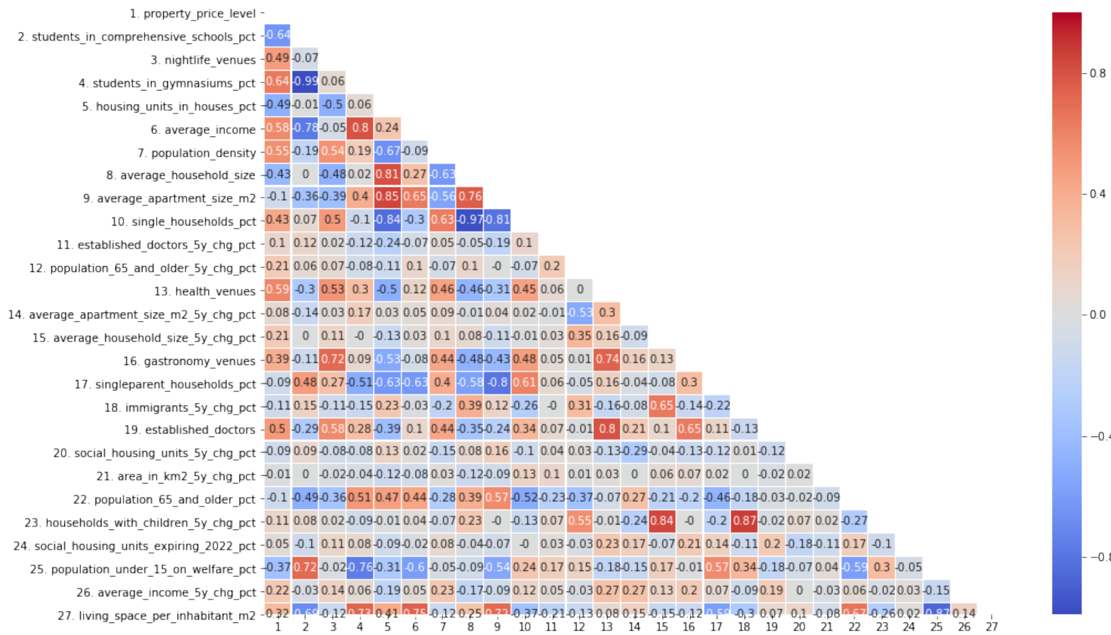
From the model we can extract an importance measure that I plot here in descending order, to answer question one from above.


students_in_comprehensive_schools_pct
nightlife_venues
students_in_gymnasiums_pct
housing_units_in_houses_pct
average_income
population_density
average_household_size
average_apartment_size_m2
single_households_pct
established_doctors_5y_chg_pct
population_65_and_older_5y_chg_pct
health_venues
average_apartment_size_m2_5y_chg_pct
average_household_size_5y_chg_pct
gastronomy_venues
singleparent_households_pct
immigrants_5y_chg_pct
established_doctors
social_housing_units_5y_chg_pct
area_in_km2_5y_chg_pct
population_65_and_older_pct
households_with_children_5y_chg_pct
social_housing_units_expiring_2022_pct
population_under_15_on_welfare_pct
average_income_5y_chg_pct
living_space_per_inhabitant_m2
employed_population_pct
social_housing_units_expiring_2021_pct
private_cars_5y_chg_pct
serviced_apartments_5y_chg_pct
students_in_secondary_schools_5y_chg_pct
households_5y_chg_pct
net_migration_5y_chg_pct
population_under_18_5y_chg_pct
dentists_5y_chg_pct
influx_from_the_environs
welfare_receivers_pct
population_under_18_migration_background_pct
social_housing_units_expiring_2019_pct
unemployed_population_5y_chg_pct
youth_unemployment_pct
residential_buildings_5y_chg_pct
public_housing_pct
youth_unemployed_5y_chg_pct
education_venues
residential_buildings
taxpayers_5y_chg_pct
living_space_per_inhabitant_m2_5y_chg_pct
housing_units_5y_chg_pct
private_cars
households_with_children_pct
housing_units_in_houses_5y_chg_pct
car_density
families_on_welfare
social_housing_units_expiring_2023_pct
influx
taxpayers
single_households_5y_chg_pct
outflux
social_housing_units_expiring_2020_pct
general_practitioners
population_under_18_migration_background_5y_chg_pct
dentists
families_on_welfare_5y_chg_pct
students_in_secondary_schools
car_density_5y_chg_pct
kindergartens_and_preschool_classes
older_unemployed_5y_chg_pct
immigrants_pct
regular_wage_earners_5y_chg_pct
population_under_15_on_welfare_5y_chg_pct
births
population_migration_background_pct
serviced_apartments
kindergartens_and_preschool_classes_5y_chg_pct
net_migration
public_transportation_venues
primary_schools
population_5y_chg_pct
deaths
singleparent_households_5y_chg_pct
influx_5y_chg_pct
general_practitioners_5y_chg_pct
welfare_receivers_5y_chg_pct
older_unemployed
births_5y_chg_pct
population_under_18_pct
housing_units
pharmacies_5y_chg_pct
pharmacies
population_migration_background_5y_chg_pct
deaths_5y_chg_pct
youth_unemployed
groceries_venues
outflux_to_environs
population_density_5y_chg_pct
population
regular_wage_earners
households
outflux_5y_chg_pct
primary_schools_5y_chg_pct
0.00
0.02
0.04
0.06
0.08
0.10
0.12

Clearly, only a few variables have a substantial influence on prices. Going down the sorted list, importances quickly drop to very low values. Interestingly, the distribution of students across the two main types of schools (comprehensive vs. Gymnasium, the most advanced type of German secondary schools) seems to be very important, followed by the number of nightlife venues, the proportion of detached and semi-detached houses vs. apartments, population density, average household size and income.

For the cluster analysis in the next section, I will only use features with an importance score of at least 0.005 (see red line).

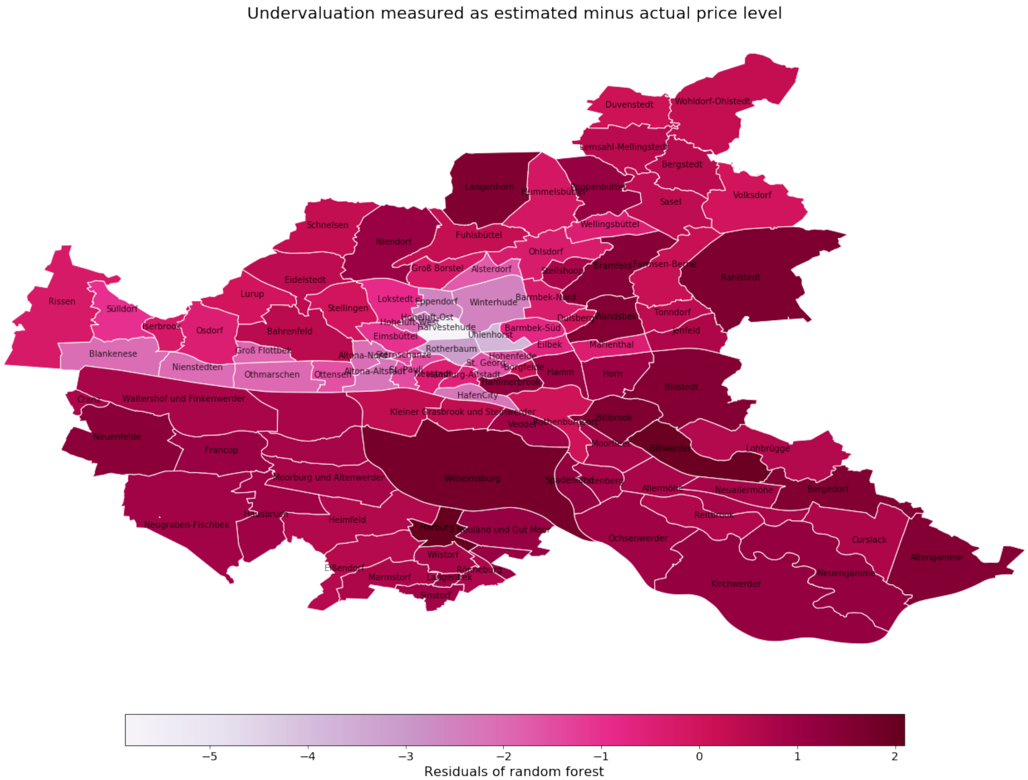
Now, to see the direction of the effects of these 26 features, let's generate a correlation matrix.



The first column shows the correlations of all neighborhood features with the price level. The highest positive correlations are percentage of students in Gymnasiums, number of health venues, and average income. The highest negative correlations are the percentage of students in comprehensive schools, the percentage of houses vs. apartments, and average household size.

4.2 Undervalued neighborhoods

To find out which neighborhoods are rather expensive and which are not in relation to their characteristics, I calculate the residuals of the random forest model, i.e. the difference between the predicted prices and the actual ones. In the plot, those neighborhoods that are undervalued according to the model (price estimate is higher than actual price) are displayed in darker shades of red.



Let's list the top five undervalued neighborhoods.

[182]:
```
                     resid  average_apartment_price_per_m2  \
name
Harburg           0.753160                          3265.0
Wellingsbüttel    0.748222                          4037.0
Wilhelmsburg      0.715014                          3249.0
Marmstorf         0.441390                          2480.0
Steilshoop        0.417820                          2635.0

                average_house_price_per_m2  average_land_price_per_m2  \
name
Harburg                             3075.0                      286.0
Wellingsbüttel                      5196.0                      856.0
Wilhelmsburg                        2810.0                      231.0
Marmstorf                           2624.0                      325.0
Steilshoop                          3112.0                      441.0

                average_apartment_price_per_m2_5y_chg_pct  \
name
Harburg                                        165.751788
Wellingsbüttel                                 135.754743
Wilhelmsburg                                   149.473935
```

```
Marmstorf                                       129.320547
Steilshoop                                      163.790494

                 average_house_price_per_m2_5y_chg_pct  \
name
Harburg                                     123.303239
Wellingsbüttel                              140.580381
Wilhelmsburg                                127.369118
Marmstorf                                   124.790988
Steilshoop                                  152.402065

                 average_land_price_per_m2_5y_chg_pct  average_price_per_m2
name
Harburg                                    109.852713           2208.666667
Wellingsbüttel                             134.657686           3363.000000
Wilhelmsburg                               159.416667           2096.666667
Marmstorf                                  105.209150           1809.666667
Steilshoop                                 121.160665           2062.666667
```

Looking at the 5 year change in average prices, all of these five neighborhoods have appreciated by at least ~130 %. This answers question two, the most undervalued neighborhoods are Harburg, Wellingsbüttel, Wilhelmsburg, Marmsdorf, and Steilshoop.

4.3 Neighborhood clusters

4.3.1 Cluster analysis

Finally, to answer question number 3, let's find out which neighborhoods are similar and how they can be grouped into clusters. Since some of the 26 features are highly correlated, I will deploy PCA once more in order to produce orthogonal variables. This time I will, however, not set the number of principal components to one, but let it be determined by the method of Minka 1 as implemented by sklearn, in order not to miss out on relevant variance.

[89]:

```
            0       1       2       3       4       5       6       7       8  \
count  99.000  99.000  99.000  99.000  99.000  99.000  99.000  99.000  99.000
mean    0.000  -0.000  -0.000  -0.000  -0.000   0.000  -0.000  -0.000   0.000
std     2.661   2.261   1.761   1.304   1.110   1.076   1.050   1.002   0.904
min    -5.307  -5.839  -4.155  -4.394  -3.220  -3.336  -2.560  -2.459  -2.268
25%    -2.290  -1.343  -0.829  -0.749  -0.650  -0.550  -0.465  -0.476  -0.554
50%    -0.121   0.138  -0.324   0.016  -0.081   0.044  -0.086  -0.051  -0.027
75%     1.865   1.286   0.312   0.561   0.530   0.465   0.450   0.457   0.571
max     6.149   9.402  11.566   6.992   3.487   6.423   5.707   3.100   2.316

            9  …      14      15      16      17      18      19      20  \
count  99.000  …  99.000  99.000  99.000  99.000  99.000  99.000  99.000
mean   -0.000  …  -0.000  -0.000   0.000  -0.000   0.000   0.000  -0.000
std     0.875  …   0.558   0.514   0.465   0.437   0.426   0.329   0.254
```

```
min    -2.438  …  -1.714  -1.548  -1.214  -0.933  -1.425  -1.001  -0.430
25%    -0.358  …  -0.294  -0.269  -0.293  -0.292  -0.236  -0.138  -0.161
50%    -0.048  …   0.034   0.034   0.024  -0.082   0.013  -0.002  -0.043
75%     0.393  …   0.245   0.343   0.288   0.262   0.247   0.147   0.108
max     3.831  …   2.051   1.594   1.343   1.261   1.868   1.014   1.050

           21      22      23
count  99.000  99.000  99.000
mean   -0.000  -0.000   0.000
std     0.236   0.153   0.102
min    -0.690  -0.378  -0.268
25%    -0.157  -0.087  -0.064
50%     0.007  -0.005  -0.010
75%     0.154   0.083   0.070
max     0.705   0.450   0.433

[8 rows x 24 columns]
```

With the resulting 24 principal components, I will run the KMeans clustering algorithm. Since, the elbow criterion didn't really work here (it resulted in a number of clusters of >100), I restrict the number of clusters to 6. This means that the clusters will not be as homogeneous as they could be, but this number is manageable for interpretations.

4.3.2 Discussion of the results

Let's now look at each of the resulting clusters and see what features are most different from the average of all neighborhood. I restrict the analysis to the top and bottom 5 features per cluster. I also add the mean of the average property price mer m² to the top of the table. Apart of the cluster mean, the table includes the difference to the grand mean of all neighborhoods in absolute terms and measured in standard deviations (*Scaled Diff.*). I'll try to be creative and give each cluster a suggestive name based on the results.

```
Neighborhoods: Hamburg-Altstadt, Hamm, Neustadt, St. Georg, Altona-Altstadt,
Altona-Nord, Bahrenfeld, Sternschanze, Lokstedt, Stellingen, Barmbek-Nord,
Barmbek-Süd, Dulsberg, Hohenfelde, Langenhorn, Bramfeld, Eilbek, Rahlstedt,
Wandsbek, Bergedorf, Harburg
```

[81]:

	Cluster Mean	Diff. Grand Mean	Scaled Diff.
average_price_per_m2	3181.39	300.76	0.26
groceries_venues	15.33	6.90	0.91
single_households_pct	61.54	10.82	0.91
gastronomy_venues	93.48	44.22	0.89
public_transportation_venues	23.95	10.65	0.88
health_venues	29.62	14.36	0.83
population_under_18_pct	13.85	-3.00	-0.77
housing_units_in_houses_pct	8.98	-22.38	-0.83
households_with_children_pct	14.48	-4.60	-0.84

average_apartment_size_m2	67.04	-17.22	-0.87
average_household_size	1.64	-0.24	-0.92

The first cluster has many of the rather central, very popular neighborhoods with a high density of restaurants, shops, and doctors. It is dominated by 62% of single households and smaller apartments with 67 mliving area [2] on average.

I'll name this cluster: *Single's paradise*

`Neighborhoods: Iserbrook, Bergstedt, Allermöhe, Altengamme, Billwerder, Curslack, Kirchwerder, Moorfleet, Neuengamme, Ochsenwerder, Reitbrook, Spadenland, Tatenberg, Francop, Langenbek, Neuenfelde, Rönneburg, Sinstorf, Moorburg und Altenwerder, Neuland und Gut Moor`

[92]:

	Cluster Mean	Diff. Grand Mean	Scaled Diff.
average_price_per_m2	1938.19	-942.44	-0.82
housing_units_in_houses_pct	68.31	36.95	1.36
car_density	477.20	115.78	1.02
average_household_size	2.14	0.26	0.99
average_apartment_size_m2	102.71	18.45	0.93
households_with_children_pct	23.18	4.11	0.75
health_venues	0.90	-14.36	-0.83
gastronomy_venues	5.80	-43.45	-0.87
public_transportation_venues	2.65	-10.65	-0.88
groceries_venues	1.00	-7.43	-0.98
single_households_pct	38.79	-11.94	-1.00

The second cluster is a lot cheaper than the average and includes more detached and semi-detached houses than apartments. It has a high car density, relatively many families with children and a low number of stores, restaurants or public transportation stops.

I'll name this cluster: *Low infrastructure Family*

`Neighborhoods: Billbrook, Billstedt, Borgfelde, Horn, Rothenburgsort, Veddel, Wilhelmsburg, Lurup, Osdorf, Eidelstedt, Schnelsen, Ohlsdorf, Farmsen-Berne, Hummelsbüttel, Jenfeld, Steilshoop, Tonndorf, Lohbrügge, Neuallermöhe, Cranz, Eißendorf, Hausbruch, Heimfeld, Neugraben-Fischbek, Wilstorf, Kleiner Grasbrook und Steinwerder, Waltershof und Finkenwerder`

[83]:

	Cluster Mean	Diff. Grand Mean \
average_price_per_m2	2174.83	-705.80
population_under_18_migration_background_pct	66.19	18.70
population_migration_background_pct	46.98	13.84
population_under_15_on_welfare_pct	31.20	12.54
welfare_receivers_pct	15.78	6.10
students_in_district_schools_pct	60.97	10.86
average_apartment_size_m2	73.70	-10.56
average_income	28281.44	-13886.09
living_space_per_inhabitant_m2	33.88	-5.95

```
students_in_secondary_schools_pct                  34.72           -12.16

                                              Scaled Diff.
average_price_per_m2                                 -0.62
population_under_18_migration_background_pct          0.93
population_migration_background_pct                   0.92
population_under_15_on_welfare_pct                    0.87
welfare_receivers_pct                                 0.76
students_in_district_schools_pct                      0.70
average_apartment_size_m2                            -0.53
average_income                                       -0.64
living_space_per_inhabitant_m2                       -0.72
students_in_secondary_schools_pct                    -0.75
```

The third cluster consists of neighborhoods with high proportion of immigrants and welfare receivers and has a low average household income.

I'll name this cluster: *Immogrants and welfare*

`Neighborhoods: HafenCity, Hammerbrook`

[84]:

```
                                          Cluster Mean  Diff. Grand Mean  \
average_price_per_m2                           4127.96          1247.33
residential_buildings_5y_chg_pct                132.73            30.68
housing_units_5y_chg_pct                        158.54            55.07
students_in_secondary_schools_5y_chg_pct        406.24           296.75
population_65_and_older_5y_chg_pct              151.16            48.24
taxpayers_5y_chg_pct                            130.31            29.12
car_density                                     195.43          -165.99
population_65_and_older_pct                       6.28           -11.64
living_space_per_inhabitant_m2_5y_chg_pct        78.97           -17.51
car_density_5y_chg_pct                           72.49           -24.02
average_apartment_size_m2_5y_chg_pct             92.32            -6.91

                                          Scaled Diff.
average_price_per_m2                              1.09
residential_buildings_5y_chg_pct                  6.31
housing_units_5y_chg_pct                          6.18
students_in_secondary_schools_5y_chg_pct          5.81
population_65_and_older_5y_chg_pct                5.61
taxpayers_5y_chg_pct                              5.54
car_density                                      -1.47
population_65_and_older_pct                      -2.13
living_space_per_inhabitant_m2_5y_chg_pct        -2.18
car_density_5y_chg_pct                           -2.89
average_apartment_size_m2_5y_chg_pct             -4.90
```

The fourth cluster is characterized by change: More new buildings, more students more older people, smaller apartments, less cars. It consists of only two neighborhoods, one of which is the

HafenCity, Hamburgs new neighborhood which was built in former industrial areas of the harbor in the middle of the city.

I'll name this cluster: *New city center*

Neighborhoods: St. Pauli, Ottensen, Eimsbüttel, Harvestehude, Hoheluft-West, Rotherbaum, Eppendorf, Hoheluft-Ost, Uhlenhorst, Winterhude

[85]:

	Cluster Mean	Diff. Grand Mean	Scaled Diff.
average_price_per_m2	5132.70	2252.07	1.96
nightlife_venues	78.20	58.58	1.83
established_doctors	179.05	126.38	1.63
population_density	10958.54	6722.99	1.63
health_venues	39.80	24.54	1.43
households_with_children_pct	14.29	-4.79	-0.87
housing_units_in_houses_pct	3.04	-28.32	-1.05
average_household_size	1.61	-0.28	-1.06
students_in_comprehensive_schools_pct	32.11	-17.73	-1.13
students_in_district_schools_pct	30.56	-19.55	-1.26

The fifth cluster contains the very hip and expensive areas around the Alster lake. It has the highest density of nightlife venues and doctors, and also the highest population density.

I'll name this cluster: *High life*

Neighborhoods: Blankenese, Groß Flottbek, Nienstedten, Othmarschen, Rissen, Sülldorf, Niendorf, Alsterdorf, Fuhlsbüttel, Groß Borstel, Duvenstedt, Lemsahl-Mellingstedt, Marienthal, Poppenbüttel, Sasel, Volksdorf, Wellingsbüttel, Wohldorf-Ohlstedt, Marmstorf

[86]:

	Cluster Mean	Diff. Grand Mean	\
average_price_per_m2	3226.65	346.02	
average_income	69167.74	27000.21	
students_in_secondary_schools_pct	66.79	19.91	
population_65_and_older_pct	23.97	6.05	
living_space_per_inhabitant_m2	48.77	8.94	
average_apartment_size_m2	104.63	20.37	
population_migration_background_pct	21.02	-12.12	

```
population_under_15_on_welfare_pct           6.57            -12.09
singleparent_households_pct                 18.17             -5.44
students_in_comprehensive_schools_pct       30.66            -19.18
students_in_district_schools_pct            31.03            -19.08

                                       Scaled Diff.
average_price_per_m2                          0.30
average_income                                1.25
students_in_secondary_schools_pct             1.22
population_65_and_older_pct                   1.11
living_space_per_inhabitant_m2                1.09
average_apartment_size_m2                     1.03
population_migration_background_pct          -0.80
population_under_15_on_welfare_pct           -0.84
singleparent_households_pct                  -0.89
students_in_comprehensive_schools_pct        -1.23
students_in_district_schools_pct             -1.23
```

The last cluster consist of neighborhoods with high income families with homes that are 20% larger than the average. Looking at the map (see below), these are often neighborhoods at the outskirts of the city.

I'll name this cluster: *Wealthy suburbs*

5 Conclusion

With this analysis, we have set out to help decision makers who are looking at the real estate market of the city of Hamburg, by answering the following questions: 1) What are the main drivers of real estate prices in Hamburg? 2) Which neighborhoods are undervalued, when relating their attributes to current average property prices? 3) How can neighborhoods be characterized? Can the 100+ neighborhoods be clustered by similarity into a manageable number of groups, in order to get a quick understanding of what they are like?

In summary,these were the answers we found by employing a set of sophisticated machine learning methods on a comprehensive set of data:

1) The main drivers of real estate prices in Hamburg are:

- Proportion of students in comprehensive schools (-) vs. in Gymnasiums (+)
- The number of nightlife venues (+)
- The proportion of detached and semi-detached houses (-) vs. apartments (+)
- The population density (+)
- Average household size (-)
- Household income (+)

2) The most undervalued neighborhoods are: Harburg, Wellingsbüttel, Wilhelmsburg, Marmsdorf, and Steilshoop.

3) The cluster analysis resulted in 6 clusters that I named:

- Singles paradise
- Low infrastructure family
- Immigrants and welfare
- New city center
- High life
- Wealthy suburbs

Of course, this analysis is not perfect. There are factors that weren't available such as how green the neighborhoods or how much noise there is. Including data about such factors would be a sensible extension of this analysis.

To conclude, lets draw a map of the six clusters.